An Automatic Digital Audio Authentication/Forensics System

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 21, 2017; accepted February 10, 2017, date of publication February 24, 2017, date of current version March 28, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2672681

An Automatic Digital Audio Authentication/Forensics System

ZULFIQAR ALI[1], MUHAMMAD IMRAN[2], AND MANSOUR ALSULAIMAN[1]
[1]Digital Speech Processing Group, Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia
[2]College of Computer and Information Sciences, King Saud University, Riyadh 12372, Saudi Arabia

Corresponding author: M. Imran (dr.m.imran@ieee.org)

This work was supported by the Deanship of Scientific Research, King Saud University, Riyadh, Saudi Arabia, under the Research Group under Project RG-1435-051.

ABSTRACT With the continuous rise in ingenious forgery, a wide range of digital audio authentication applications are emerging as a preventive and detective control in real-world circumstances, such as forged evidence, breach of copyright protection, and unauthorized data access. To investigate and verify, this paper presents a novel automatic authentication system that differentiates between the forged and original audio. The design philosophy of the proposed system is primarily based on three psychoacoustic principles of hearing, which are implemented to simulate the human sound perception system. Moreover, the proposed system is able to classify between the audio of different environments recorded with the same microphone. To authenticate the audio and environment classification, the computed features based on the psychoacoustic principles of hearing are dangled to the Gaussian mixture model to make automatic decisions. It is worth mentioning that the proposed system authenticates an unknown speaker irrespective of the audio content i.e., independent of narrator and text. To evaluate the performance of the proposed system, audios in multi-environments are forged in such a way that a human cannot recognize them. Subjective evaluation by three human evaluators is performed to verify the quality of the generated forged audio. The proposed system provides a classification accuracy of $99.2\% \pm 2.6$. Furthermore, the obtained accuracy for the other scenarios, such as text-dependent and text-independent audio authentication, is 100% by using the proposed system.

INDEX TERMS Digital audio authentication, audio forensics, forgery, machine learning algorithm, human psychoacoustic principles.

I. INTRODUCTION

With the recent unprecedented proliferation of smart devices such as mobile phones and advancements in various technologies (e.g., mobile and wireless networks), digital multimedia is becoming an indispensable part of our lives and the fabric of our society. For example, unauthentic and forged multimedia can influence the decisions of courts as it is admissible evidence. With continuous advancements in ingenious forgery, the authentication of digital multimedia (i.e., image, audio and video) [1] is an emerging challenge. Despite reasonable advancements in image [2], [3] and video [4], digital audio authentication is still in its infancy. Digital authentication and forensics involve the verification and investigation of an audio to determine its originality (i.e., detect forgery, if any) and have a wide range of applications [5]. For example, the voice recording of an authorized user can be replayed or manipulated to gain access to secret data. Moreover, it can be used for copyright applications such as to detect fake MP3 audio [6].

Audio forgery can be accomplished by copy-move [7], deletion, insertion, substitution and splicing [8], [9]. The applications of copy-move forgery are limited compared with other methods as it involves moving a part of the audio at other location in the same liaison. On the other hand, the deletion, insertion, substitution and splicing of forged audio may involve merging recordings of different devices, speakers and environments. This paper deals with a splicing forgery (i.e., insertion of one or more segments to the end or middle), which is more challenging. The primary objective of the proposed system is to address the following issues with high accuracy and a good classification rate:

- Differentiate between original and tampered audio generated by splicing recordings with the same microphone and different environments.

- Environment classification of original and forged audio generated through splicing. Identify forged audio irrespective of content (i.e., text) and speaker.
- Reliable authentication with forged audio of a very short duration (i.e., ∼5 seconds).

In the past, audio authentication has been achieved by applying various algorithms [5], [10]–[12]. One of the basic approaches is the visual investigation of the waveform of an audio to identify irregularities and discontinuities [12]. For example, the analysis of spectrograms [12] may reveal irregularities in the frequency component during the investigation. Similarly, listening to audio [11] may also disclose abrupt changes and the appearance of unusual noise. These methods may help to decide whether the audio is original or tampered. However, one of the prime limitations of These approaches is that they are human-dependent, where judgement errors cannot be ignored. Moreover, the availability of sophisticated manipulation tools [13], [14] makes it convenient to manipulate audio without introducing any abnormalities. Consequently, it becomes very difficult to identify those abnormalities. For example, the visual inspection of the waveform and spectrogram of the tampered audio depicted in Fig. 1 does not provide any clue of irregularity and hearing is also quite normal.

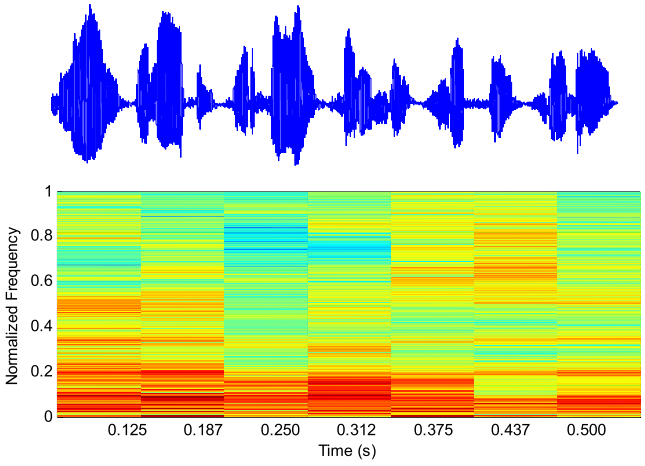

FIGURE 1. A tampered audio with its spectrogram.

To avoid human involvement, Kraetzer et al. [15] suggested an automatic system based on a machine learning algorithm. The authors claimed it as a first practical approach towards digital audio forensics that classifies microphones and the environment. Mel-frequency cepstral coefficients (MFCCs) with some time-domain features were extracted from audio for authentication. The authors in [16] also performed environment classification by using MFCCs and MPEG-7. However, the obtained accuracy was only approximately 95%. Electric network frequency has also been used in many studies for the authentication of digital audio [17]–[19]. Moreover, modified discreet cosine transformation was used in [6] for the authentication of compressed audio.

Recently, the authors in [20] used measures such as ambient noise with the magnitude of the impulse response of an acoustical channel for source authentication and the detection of audio splicing. To evaluate the method, TIMIT and another database developed in four different environments were used. Samples of 30 seconds were generated for the testing purpose. However, in real life, it is either difficult or impractical to obtain audio of such a duration for authentication. The Gaussian mixture model (GMM) has been used as a classification technique, and the obtained false-positive rate is greater than 3%. Another recent work [21] used discrete wavelet packet decomposition to identify forgery in audio. Audio samples recorded at different frequencies were used to test the system. However, the obtained accuracy is lower than [20] for the detection of normal (i.e., 86.89%) and forged audio (i.e., 89.50%). Moreover, the improper adjustment of five different parameters may increase false alarm and false rejection, which ultimately affect the accuracy of the system.

To deal with splicing forgery, this paper presents a novel audio authentication system based on human psychoacoustic (AAHP) principles of hearing. By using recordings by the same microphone but in different environments, we develop a database of normal and splicing-based forged audio containing digits from zero to nine. Forged recordings are developed by merging the digits of two different recordings after calculating their endpoints. Various measures such as total amplitude, zero crossing (ZC) and the duration of a digit are considered to determine endpoints accurately. On the other hand, digit clipping is used to generate normal recording without any modification. The three psychoacoustic principles of hearing (i.e., critical bandwidth, equal-loudness curve and cube root compression) are used to simulate the human perception of sound. The features in the proposed system are extracted by applying the hearing principles sequentially on the spectrum of audio. The features are computed from each audio and provided to GMM [22], [23] for the generation of acoustic models for the original and forged audio during the training phase of the proposed system. The generated models are then used for audio authentication and environment classification. The quality of the generated forged audio is validated by three human evaluators. The performance evaluation confirms the effectiveness and efficiency of the proposed system. The proposed system achieves a classification accuracy of 99.2%±2.6 and 100% in some cases. To the best of our knowledge, this is the first automatic audio authentication system based on hearing principles that can classify audio from the same microphone (intra-microphone authentication), but different recording environments (inter-environment) as well as an unknown speaker (speaker-independent) and known (text-dependent) and unknown text (text-independent).

The rest of the paper is organized as follows. Section 2 describes the proposed automatic audio authentication system and the generation of forged audio as well as the process for the accurate calculation of the endpoints.

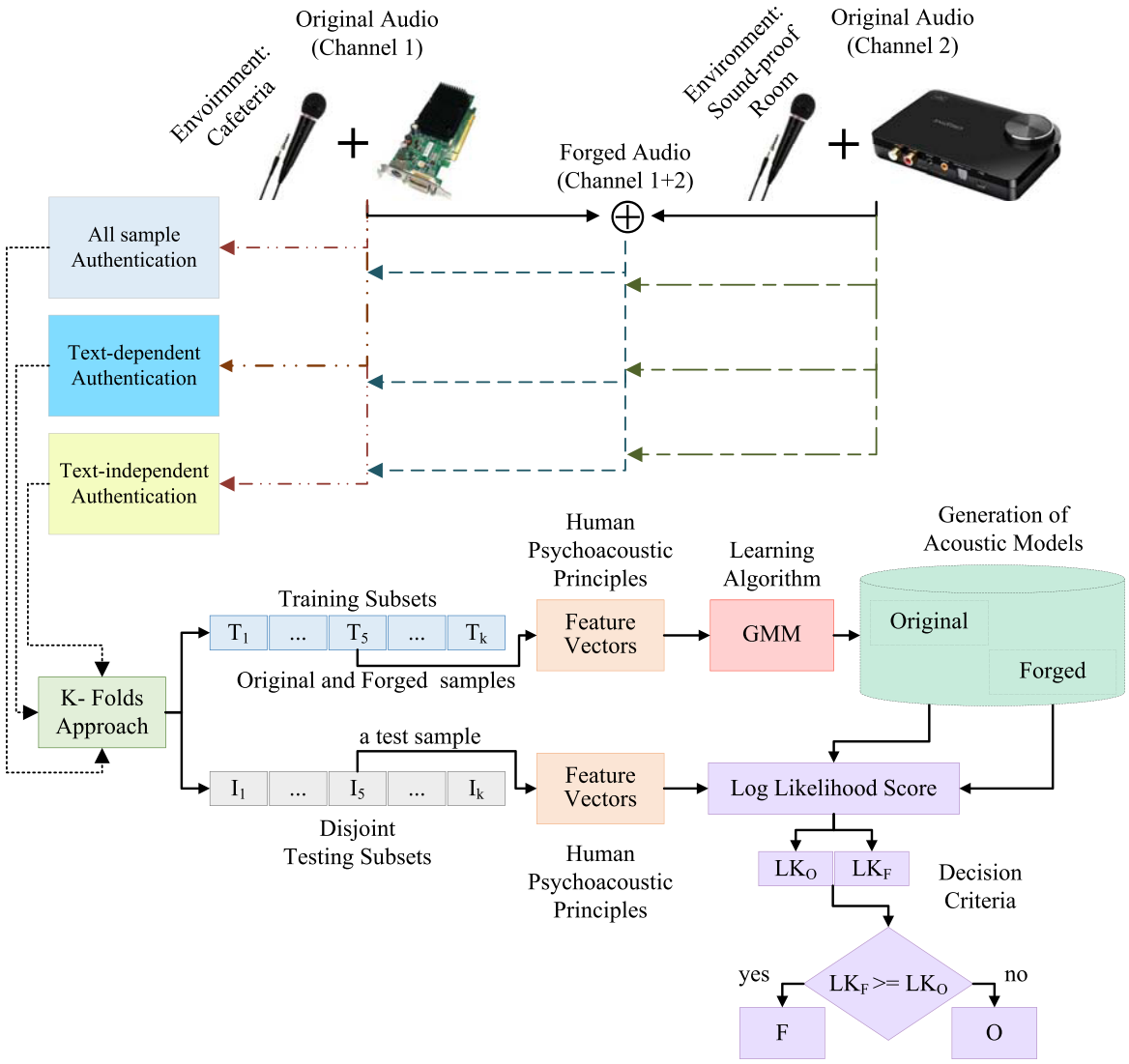

FIGURE 2. Block diagram of the proposed audio authentication system.

The subjective evaluation of the generated forged audio by the three human judges and experimental results of the proposed system are provided in Section 3. Section 4 provides the necessary discussion and compares the proposed system with some recent studies. Finally, concluding remarks and future research directions are indicated in Section 5.

II. DEVELOPMENT OF FORGED CORPUS AND AUTHENTICATION SYSTEM

This section mainly consists of two parts. The first part describes the process of splicing-based forged audio database development, clipping of normal audio and endpoint detection. The second part elucidates the robustness of the proposed system against recording text and speakers, k-folds cross validation, feature extraction based on the psychoacoustic principles of human hearing and a machine learning algorithm. Fig. 2 depicts a block diagram of the proposed automatic audio authentication system. The components of the system are described in the following subsections.

A. GENERATION OF FORGED AUDIO CORPUS

The generation of tampered audio, in a way that a human evaluator cannot guess it is so, is a big challenge and one of the crucial steps towards the development of the proposed system. Forged and normal audio samples are generated by using the King Saud University Arabic Speech Database (KSU-ASD) [24]. The reason for selecting the KSU-ASD is its diversity in recorded text, recording environments and equipment [25], [26]. To the best of our knowledge, none of the existing publicly available databases serves our purpose. The KSU-ASD is publicly available through the Linguistic Data Consortium, which is hosted by the University of Pennsylvania, Philadelphia, USA. Although the language of the KSU-ASD is Arabic, the proposed system will work for any language.

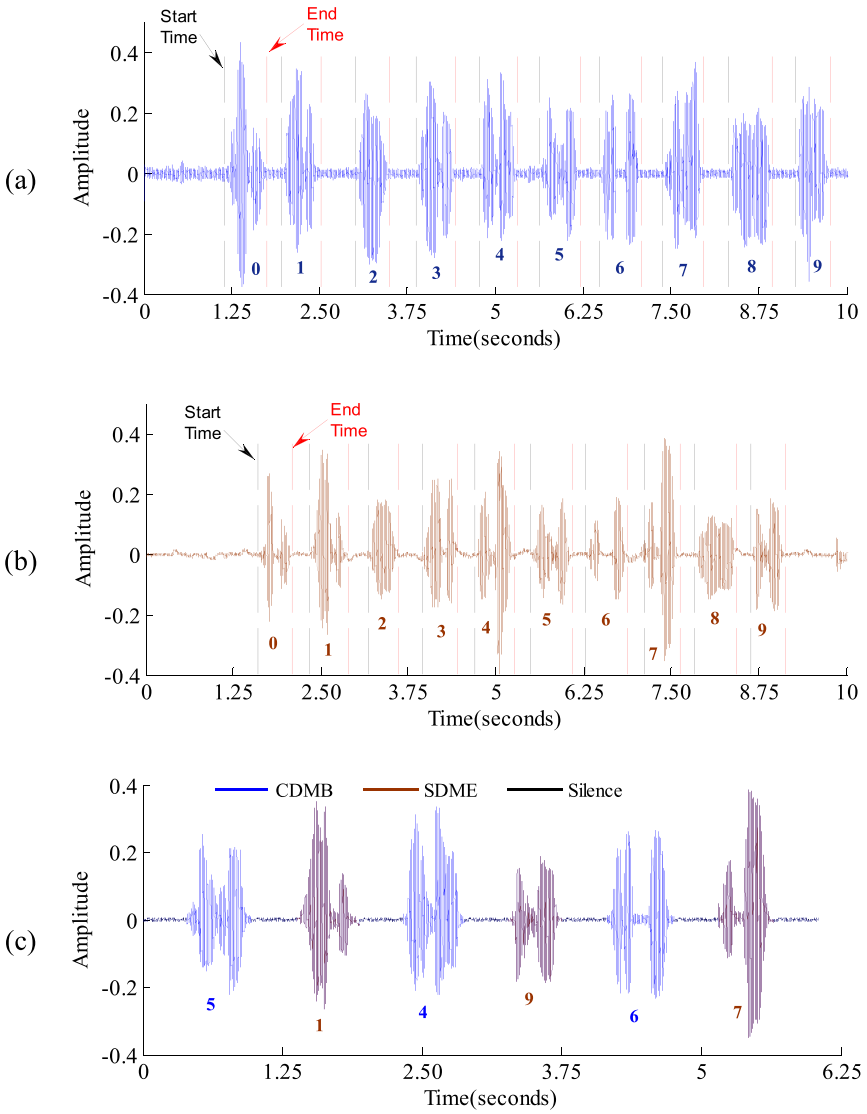

FIGURE 3. (a) The endpoints of each digit in an audio of CDMB (0–9) (b) The endpoints of each digit in an audio of SDME (0–9) (c) The resultant forged audio.

1) GENERATION OF FORGED AUDIO BY SPLICING

The KSU speech database was recorded in three different environments i.e., office (normal), cafeteria (noisy) and sound-proof room (quiet). In this study, two very different environments, cafeteria and sound-proof room, are mixed to generate the forged audio. Mixing the sound-proof room with the cafeteria is the worst case scenario, where the former represents an absolutely quiet environment and the latter represents a noisy environment containing background noise. Audio is forged by mixing the speech of two different recording settings:

1. Recording of digits in the cafeteria with a microphone (Sony F-V220) attached to a built-in sound card on the desktop (OptiPlex 760) through an audio-in jack. This is denoted as CDMB (Cafeteria, Digits, Microphone, Built-in sound card).
2. Recording of digits in the sound-proof room with a microphone (Sony F-V220) connected to an external sound card (Sound Blaster X-Fi Surround 5.1 Pro) through the USB port of the desktop (OptiPlex 760). This is represented by SDME (Sound-proof room, Digits, Microphone, External sound card).

Although it is ideal to forge an audio recording through a mobile phone because a person is unaware of the recording in such a scenario, his/her speech can be used for any purpose. However, the amplitude of mobile phone recordings is low in the KSU-ASD compared with the microphones, and through visualization, it is easy to identify that the audio is forged. Therefore, the recording of the mobile phone is not used to generate forged samples.

Fig. 3 describes the process used to generate forged audio by using the recordings of CDMB and SDME. The whole process of generating forged audio is automatic, and the first step is the generation of six-digit unique random numbers such as 514967. The range of each digit is a number from one to nine. The second step is the calculation of the endpoints of

digits in the recordings of CDMB and SDME. The process to extract the endpoints is explained in Section 2.1.2. The calculated starting and ending points of each digit in the audio of CDMB and SDME are shown in Fig. 3 (a) and 3 (b), respectively. The vertical black and red dotted lines represent the starting and ending times, respectively. The endpoints will be used to extract the digits from the recordings and mixed together for the generation of forged audio.

Once the endpoints are calculated, the odd and even digits of the random number are taken from CDMB and SDME, respectively. For example, in the random number 514967, the digits 5, 4 and 6 belong to CDMB and the remainder to SDME. In the last step, the extracted digits are combined and the resultant forged audio is depicted in Fig. 3 (c). The audio samples of CDMB and SDME, shown in Fig. 3, are recorded by speaker 1 (NS1) in the KSU-ASD database.

By using these two audios of speaker 1, eight different forged samples are generated. One of the eight forged signals is shown in Fig, 3 (c), while the remaining seven are 347268, 243157, 962351, 123456, 234567, 345678 and 456789. Forged audio containing random digits is denoted by CS_{Rand1}, CS_{Rand2}, CS_{Rand3} and CS_{Rand4}, while continuous digits are represented by $CS_{Cont1:6}$, $CS_{Cont2:7}$, $CS_{Cont3:8}$ and $CS_{Cont4:9}$. Moreover, the four original audios of CDMB are clipped in the following four different ways: 123456, 234567, 345678 and 456789; these are represented by $C_{Cont1:6}$, $C_{Cont2:7}$, $C_{Cont3:8}$ and $C_{Cont4:9}$. Similarly, the four original audios of SDME are clipped and denoted as $S_{Cont1:6}$, $S_{Cont2:7}$, $S_{Cont3:8}$ and $CS_{Cont4:9}$. Here, clipping refers to cutting each digit from the number without any modification. In this way, from the two samples of speaker 1, eight forged and eight original samples are produced. In other words, 16 samples are produced from two utterances of a speaker. In this study, 90 different speakers are considered; hence, we have 720 ($= 8 \times 90$) forged audio recordings and 720 ($= 8 \times 90$) original recordings. The total number of audio recordings in the data set is 1440.

Without repetition, 60480 unique random numbers can be generated by using one to nine digits, i.e., 9! / (9 - 6)!. Although the same number of forged audios can be generated from the two utterances, only eight tampered audio samples are produced to keep the balance between the original and forged recordings. As there are only four different possible ways to clip original audio, a maximum of four audios can be produced from an utterance. For a speaker, there are two utterances (one in each environment); therefore, eight audio recordings are possible at most.

2) PROCESS FOR ENDPOINT DETECTION

Endpoint detection is a key process in the generation of tampered audio. If digits are not extracted properly from audio samples, then their mixing will not be flawless, and hence this may mislead a human judge to wrongly perceive the audio as a tampered sample. In such a case, when an audio can be judged by listening or visualizing, then there is no purpose to build an automatic authentication system. This is the reason that forged audio is generated sophisticatedly so that nobody may guess its type, i.e., original or tampered. Therefore, different measures are used for the accurate extraction of the endpoints.

Before applying the various measures to detect the endpoints, an audio is divided into short frames. As stated earlier, speech varies quickly with respect to time, which makes it difficult to analyze. A frame of 20 milliseconds is used to compute the measures. The size of the frame is kept small to exclude it if it contains silence. In this way, the exact starting time of a digit can be determined. One of the computed measures for endpoint detection is the total amplitude, T_{amp}, of a frame, and this is given by Eq. (1) as

$$T^{i}_{amp} = \sum_{i=1}^{n} |a_i| \tag{1}$$

where $[a_1, a_2, a_3, \ldots, a_n]$ are the corresponding amplitudes for the samples $[x_1, x_2, x_3, \ldots, x_n]$ in the i^{th} frame X_i of the audio signal $X = [X_1, X_2, X_3, \ldots, X_N]^T$. The signal is divided into N non-overlapping frames and the number of samples in each frame is 400, i.e., $n = 400$. The length of each frame is 20 milliseconds and each audio is down-sampled to 16 KHz. A threshold to detect the silence frames and voiced parts of the audio is shown by a horizontal line in Fig. 4 (a) and is given by Eq. (2) as

$$thresh = 3\% \text{ of } \left[\max\left(T_{amp}\right) - \min\left(T_{amp}\right)\right] + \min\left(T_{amp}\right) \tag{2}$$

The other measure for the calculation of the endpoints is ZC. In the case of silence, the amplitude in an audio should be zero, but this is not the case. Due to background noise during silence, an audio contains low amplitude and ZC is high. To make the ZC equal to zero for the silence part, four times the maximum absolute amplitude in a frame, to a minimum T_{amp}, is subtracted from the whole audio. By doing so, the amplitude for silence in the whole audio becomes negative and no ZC will be there. It can be observed in Fig. 4 (a) that ZC for the silence part is zero now. Moreover, a threshold equal to 2% of the maximum ZC is also adjusted and ZC below this represents a silence part. These parameters for ZC are adjusted by investigating different audio recordings.

During the phonation of some digits, some speakers give a short pause. For example, in the case of the Arabic digit 6 (*sit-tah*), speakers pronounce it as *sit*-short pause-*tah*. Therefore, the waveform of a digit is split into two parts, as shown in Fig. 4 (b). To handle this situation, a check is implemented on the duration of a digit. The normal duration of a digit is ~ 0.5 seconds and the silence between the digits is ~ 0.4 seconds. If the duration of each of the two consecutives parts of the speech is less than 0.3 seconds, then it means that a digit is split into two parts. In addition, a silence of less than 0.3 seconds between the consecutive parts also confirms the situation. These conditions are implemented by

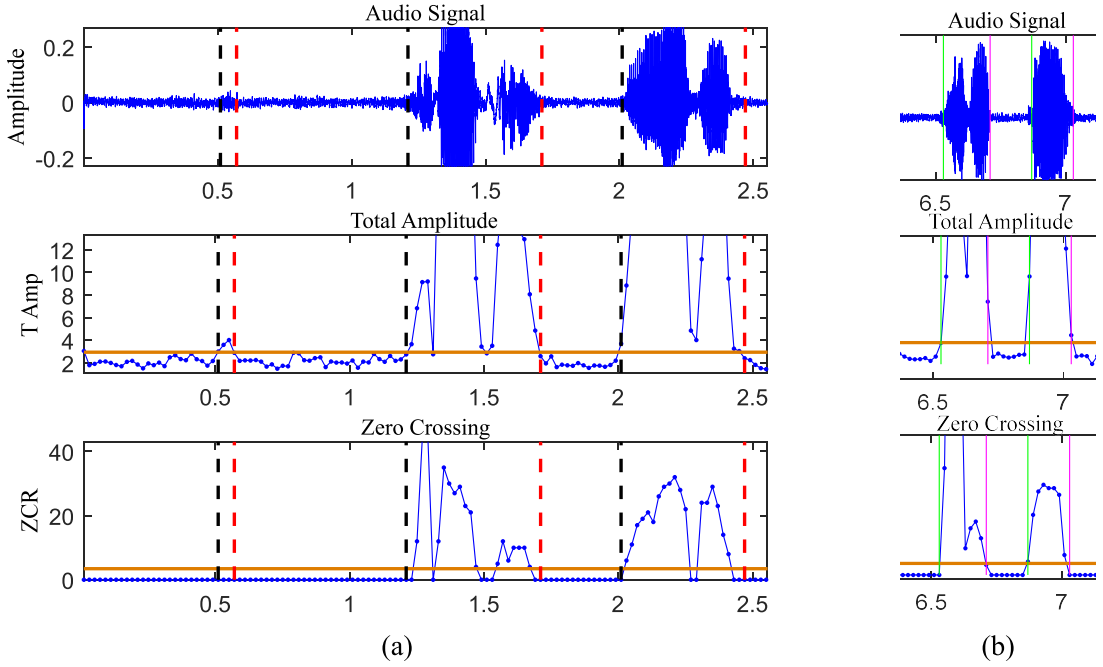

FIGURE 4. (a) Accurate calculation of the endpoints of digits by using T_{amp} and ZC (b) A digit is split into two parts.

using Eq. (3):

$$\begin{aligned} &\textit{if } [\text{duration}(SegX, SegY)\text{AND silence } (SegX, SegY)] < 0.3 \\ &\textit{then } \text{merge } (SegX, SegY) \end{aligned} \tag{3}$$

where $SegX$ and $SegY$ are the two split parts of a digit. Finally, the starting time for such digits will be the starting time of the first split part and the ending time will be the ending time of the second split part.

B. PROPOSED AUTOMATIC AUDIO AUTHENTICATION SYSTEM

The major components of the proposed automatic authentication system are described in this section. The system is evaluated by using distinctive text and a set of speakers for training and testing to make the system robust against text and speakers. Through the cross validation approach, the system is also evaluated by using each recording of the developed forged database. Three principles of the human hearing system are used to extract the features, which are added into the GMM for the automatic authentication of the audio and environment classification.

1) TEXT ROBUSTNESS, SPEAKER INDEPENDENCE AND CROSS VALIDATION

To observe the robustness of the proposed authentication system against recorded text, two types of experiments are performed. The experiments in which the same text is used to train and test the system are referred to as text-dependent authentication, while those experiments in which the system is trained and tested with distinct text are referred to as text-independent experiments. Moreover, in all experiments, the speakers used to train and test the system are different from each other. This means that the system can authenticate the audio of an unknown person. In addition, the proposed system is tested with each sample by using the k-folds cross validation approach to avoid bias in the training and testing data. In k-folds cross validation, the whole data set of the original and forged audio is divided into k-disjoint subsets. Each time one of the subsets is used for testing, the remaining $k-1$ are used for training.

2) FEATURE EXTRACTION

Feature extraction from the original and tampered audio is one of the key components of the proposed system. The features are extracted by applying the psychoacoustic principles of human hearing [27]. A set of three human psychoacoustic principles, namely the critical band spectral estimation, equal loudness hearing curve and intensity loudness power law of hearing, are implemented to compute the feature vectors for the proposed system. The audio of a person varies quickly over time, which makes it difficult to analyze. Therefore, before applying the psychoacoustic principles, the audio is split into very small blocks. In each block, the behavior of the speech is quasi-stationary, and hence can be analyzed easily. To avoid the loss of information at the ends, a new block is overlapped with the previous by 50%. Moreover, to ensure the continuity of the audio in successive blocks, it is necessary to taper the ends of the divided blocks to zero. The blocks are tapered by multiplying by the hamming window [28], [29], which is given by Eq. (4):

$$h_w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad \text{where } 0 \le n \le N-1 \tag{4}$$

where N is the fixed length of the blocks. The hamming window is multiplied by each block of the audio signal $A(n)$ and the resultant signal is represented by A_h. The multiplication of the hamming window h_w by the i^{th} block of the audio signal $A(n)$ is given as:

$$A_{h_i}(n) = A_i(n) \times h_w(n) \tag{5}$$

In this way, spectral leakage can also be avoided when Fourier transformation (FT) is applied. FT is an important component in the computation of the feature vectors, which transform a signal from the time domain to the frequency domain and provide energy information at each frequency component. The operation of convolution in the time domain becomes a simple multiplication in the frequency domain, which makes the rest of the calculation easier. The output obtained after applying FT is referred to as a spectrum of the input audio. The obtained spectrum of the windowed-audio signal A_h is given in Eq. (6):

$$\begin{aligned} A_k^s &= F\{A_h\} \\ A_k^s &= \sum_{n=0}^{N-1} A_h(n) \times e^{\frac{-2\pi jkn}{N}} \end{aligned} \tag{6}$$

where F stands for FT.

The further analysis of the spectrum is done by applying the different psychoacoustic conditions of hearing to obtain the feature vectors. During auditory perception, human ears respond differently to different frequencies. The role of human ears is vital in separating the frequencies, and they transmit them to the basilar membrane (BM). The lower frequencies are localized towards the apex, while the higher are confined at the basal turn. Each location on the BM acts like a band-pass filter. Moreover, the positioning of the resonant frequencies (bandwidth of frequencies) along the BM is linear up to 500 Hz and logarithmic above it. The distribution of the frequency along the BM can be approximated by using Eq. (7):

$$Bark = 13 \arctan(0.00076\, f) + 3.5 \arctan\left(\frac{f}{7500}\right)^2 \tag{7}$$

where f is frequency in Hz and one bark represents one critical band. The relation was proposed by Zwicker [30]. Twenty-four bark-spaced filters are used in the study and they correspond to the first 24 critical bands of hearing. After applying the bark scale on spectrum A^s, bark-wrapped spectrum A_B is given by Eq. (8):

$$A_B\,(p, Fr) = Bark\,(p, b) \times A^s\,(b, Fr) \tag{8}$$

where p and b stand for the number of filters and FT bins, respectively, Fr denotes the number of frames (blocks) in the audio signal and A^s represents the spectrum of the windowed-audio signal.

The bark-warped critical band spectrum A_B is now passed through a relative spectra band-pass filter to remove the effect of the constant and slowly varying parts in each component of the estimated critical band spectrum [31]. The human auditory system is relatively insensitive to those slowly varying stimuli. The response of the filter is given by Eq. (9):

$$R(z) = z^4 \times \frac{\left(0.2 + 0.1z^{-1} - 0.1z^{-3} - 0.2z^{-4}\right)}{1 - 0.94z^{-1}} \tag{9}$$

The output spectrum is denoted by A_R. The study of physiological acoustics shows that the sensitivity of human auditory mechanisms to different frequencies is different at the same sound intensity. To incorporate the phenomenon that human hearing is more sensitive to the middle frequency range of the audible spectrum, A_R is multiplied by an equal loudness curve to approximate the equal loudness of human hearing at different frequencies. The equal loudness weight for the j^{th} filter E_j of critical band spectrum A_R is calculated as

$$E_j = \frac{f_j^2 \times \left(f_j^2 + 1.44 \times 10^6\right)}{\left(f_j^2 + 1.6 \times 10^5\right) \times \left(f_j^2 + 9.61 \times 10^6\right)} \tag{10}$$

The center frequency of the j^{th} filter is represented by f_j and the obtained spectrum is represented by A_E.

According to the power law of hearing, a nonlinear relationship exists between the intensity of sound and perceived loudness [32]. The phenomenon is incorporated after taking the cube root of the spectrum, which compresses the spectrum, and the obtained output is referred to as the processed auditory spectrum of the input audio. The auditory processed spectrum is our required feature vectors, denoted by A_C in Eq. (11), and this is obtained after the cube root as

$$A_C = \sqrt[3]{A_E} \tag{11}$$

3) AUDIO AUTHENTICATION AND ENVIRONMENT CLASSIFICATION

The feature vectors are extracted in both phases of the proposed system. In the training phase, the feature vectors are computed from the subsets of the normal and forged audio obtained after the k-folds scheme and provided to the GMM to generate acoustic models for each of them (i.e., one model for the original and the other for the forged). The GMM is state-of-the-art modeling and has been used in many scientific areas [33]–[35]. The initial parameters of the GMM are selected by using the k-means algorithm. These parameters are estimated and tuned by the well-known expectation maximization algorithm [36] to converge to a model giving a maximum log-likelihood value. In the testing phase, the feature vectors are extracted from an unknown audio and compared with the acoustic model of the original and tampered audio. The log-likelihood for each model is then computed. If the log-likelihood value is greater for the forged acoustic model, then the unknown audio is tampered; otherwise, it is an original.

Moreover, in the case of environment classification, the GMM generates one model for each environment. An unknown audio compared with each environment and

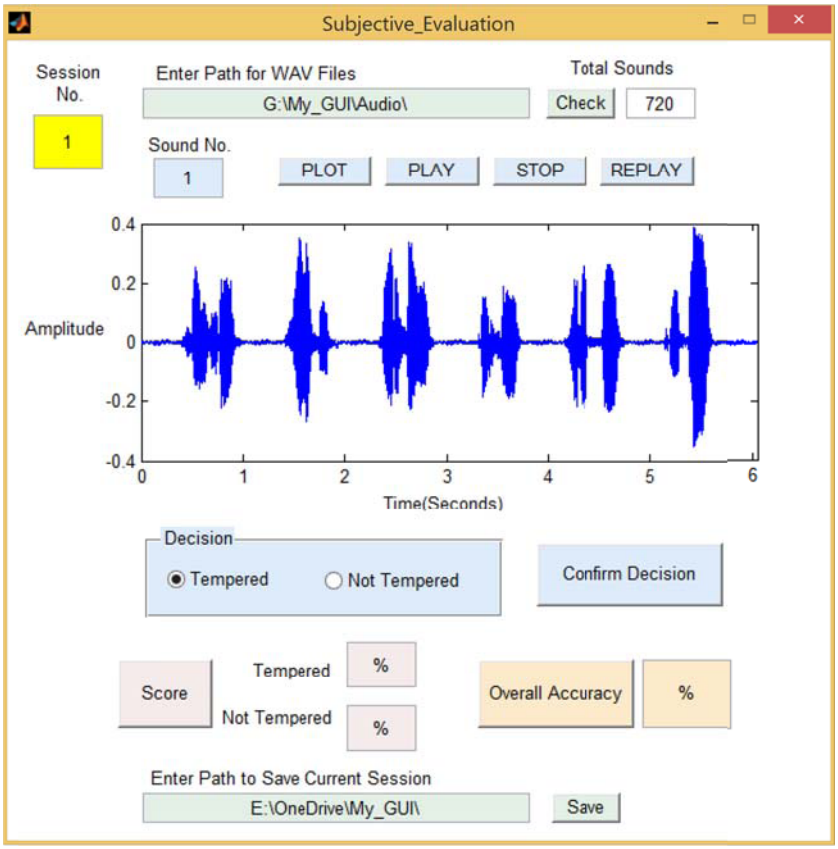

FIGURE 5. GUI for a subjective evaluation of audio.

the model having maximum log-likelihood will be the environment of that unknown audio. The following subsection presents the generation of forged audio and procedure for endpoint detection.

III. PERFORMANCE EVALUATION

To validate the performance, human evaluators and the proposed automatic audio authentication system are used. This section describes the environment setup, performance metrics, experimental results and analysis.

A. SUBJECTIVE AUDIO EVALUATION

As stated in the previous section, sophisticated forged audio recordings are generated to make it difficult to distinguish between tampered and original audio. To observe the quality of the generated tampered audios, they are evaluated by three human evaluators, named Judge 1, Judge 2 and Judge 3. For this purpose, a graphical user interface (GUI) is developed and shown in Fig. 5. All evaluators are holders of master's degrees in sciences and do not have any known visual or hearing problem. It is not necessary for a judge to evaluate all audio in one session. A judge can generate a report for each session to track the audio evaluated. The report provides the following information for a session: session X started with sound Y and ended with sound Z, the total number of evaluated audio (number of original and tampered audio) and evaluation metrics such as true positives, false positives, true negatives, false negatives and overall accuracy. These metrics are described later in this section.

The judges provide the path of audio recordings and then check how many recordings are available for the evaluation. They enter a sound number, and can plot, play, stop and replay the sound to make a decision. They have two options to evaluate an audio i.e., sound and visual. To enter the decision, they select one of the radio buttons and confirm the decision by pressing ''Confirm Decision''.

One of the most important steps in the subjective evaluation is the name of the original and tampered audio files. The audio files are provided with the developed GUI for evaluation. If the judges can guess the type of audio from its name, then the whole procedure is useless. Therefore, an 8-digit number is used to name an audio file, for instance, '24901684.wav'. If the sum of digits at even and odd places in a filename is even, then the audio is original. On the other hand, if the sum of digits in the odd places is odd and in the even places is even, then the audio is tampered. For example in '24901684.wav', the sum of digits in the odd places is even ($20 = 2 + 9 + 1 + 8$) and the sum of digits in the even places is also even ($14 = 4 + 0 + 6 + 4$); therefore, audio '24901684.wav' is original. The GUI checked this automatically to determine that the decision entered by a judge is correct or not. In a GUI, tampered audio is considered as a positive class, while original audio is treated as a negative class. The subjective

TABLE 1. Subjective evaluation by judge 1, judge 2 and judge 3.

Human Evaluator	Normal		Forged	SEN	SPE	ACC
Judge 1	$C_{Cont1:6}$	$S_{Cont1:6}$	$CS_{Cont1:6}$	51.11	47.22	48.52
	$C_{Cont2:7}$	$S_{Cont2:7}$	$CS_{Cont2:7}$	56.67	52.22	53.70
	$C_{Cont3:8}$	$S_{Cont3:8}$	$CS_{Cont3:8}$	44.44	48.33	47.04
	$C_{Cont4:9}$	$S_{Cont4:9}$	$CS_{Cont4:9}$	53.33	54.44	54.07
Judge 2	$C_{Cont1:6}$	$S_{Cont1:6}$	$CS_{Cont1:6}$	45.56	48.89	47.78
	$C_{Cont2:7}$	$S_{Cont2:7}$	$CS_{Cont2:7}$	54.44	55.56	55.19
	$C_{Cont3:8}$	$S_{Cont3:8}$	$CS_{Cont3:8}$	57.78	50.56	52.96
	$C_{Cont4:9}$	$S_{Cont4:9}$	$CS_{Cont4:9}$	52.22	50.00	50.74
Judge 3	$C_{Cont1:6}$	$S_{Cont1:6}$	$CS_{Cont1:6}$	50.00	49.44	49.63
	$C_{Cont2:7}$	$S_{Cont2:7}$	$CS_{Cont2:7}$	55.56	53.89	54.44
	$C_{Cont3:8}$	$S_{Cont3:8}$	$CS_{Cont3:8}$	43.33	53.33	50.00
	$C_{Cont4:9}$	$S_{Cont4:9}$	$CS_{Cont4:9}$	45.56	46.67	46.30

evaluation by Judge 1, Judge 2 and Judge 3 is provided in Table 1.

The results of the experiments are evaluated by using the following performance metrics: sensitivity (SEN), specificity (SPE) and accuracy (ACC). SEN is a ratio between truly detected tampered audio and the total number of tampered audios. SPE is a ratio between truly classified original audio and the total number of original audios. ACC is a ratio between truly identified audio and the total number of audios. The measures are calculated by using the following relations:

$$SEN = \frac{true\ Temp}{true\ Temp + false\ Orig} \times 100 \tag{12}$$

$$SPE = \frac{true\ \text{Orig}}{true\ Orig + false\ Temp} \times 100 \tag{13}$$

$$ACC = \frac{trure\ Temp + true\ Orig}{total\ Orig + total\ Temp} \times 100 \tag{14}$$

where *true Temp* means a tampered audio is detected as a tampered audio by the system, *false Orig* means a tampered audio is detected as an original audio, *true Orig* means an original audio is detected as an original audio, *false Temp* means an original audio is detected as a tampered audio by the system, *total Orig* represents the total number of original audios and *total Temp* stands for the total number of tampered audios.

$C_{Cont1:6}$, $C_{Cont2:7}$, $C_{Cont3:8}$ and $C_{Cont4:9}$ belong to the channel CDMB and $S_{Cont1:6}$, $S_{Cont2:7}$, $S_{Cont3:8}$ and $S_{Cont4:9}$ are taken from the channel SDME. CS_{Rand1}, CS_{Rand2}, CS_{Rand3}, CS_{Rand4}, $CS_{Cont1:6}$, $CS_{Cont2:7}$, $CS_{Cont3:8}$ and $CS_{Cont4:9}$ are the eight forged audio recordings. In the subjective evaluation, $CS_{Cont1:6}$, $CS_{Cont2:7}$, $CS_{Cont3:8}$ and $CS_{Cont4:9}$ are used only because they have the same pattern of digits in each audio as the channels CDMB and SDME have.

Each judge performs four different types of experiments. In the first experiment, the recorded text of the audio is digits 1 to 6 and the obtained accuracies are 48.52%, 47.78% and 49.63% for Judge 1, Judge 2 and Judge 3, respectively. The results are lower than 50%, which shows that the generated tampered audio is very similar to the original audio. In a two-class problem, a sample has a 50% probability for each class, but in our case, the obtained results are even less than 50%, confirming that a judge has no clue about the class of the audio (i.e., the results are random). A similar type of trend is found in the obtained accuracies of the other experiments; either accuracy is lower than 50% or just greater than 50%. In the next section, the automatic authentication of the audio is performed by using the proposed system, and the results are compared with the subjective evaluation.

B. AUTOMATIC AUDIO AUTHENTICATION THROUGH THE PROPOSED SYSTEM

Automatic audio authentication is performed by means of the proposed automatic authentication system. Various experiments are conducted by considering the different scenarios to observe the performance of the proposed system. Experiments are classified into three major categories.

In the first category, all original and forged audios of both channels are used, and the results are provided in Table 2. The results of the experiments are presented by using the same metrics described earlier (i.e., SEN, SPE and ACC),

TABLE 2. Automatic authentication results by using all normal and forged audio samples.

Normal Samples	Forged Samples	GMM	SEN±STD	SPE±STD	ACC±STD	AUC
$C_{Cont1:6}$, $C_{Cont2:7}$, $C_{Cont3:8}$, $C_{Cont4:9}$ $S_{Cont1:6}$, $S_{Cont2:7}$, $S_{Cont3:8}$, $S_{Cont4:9}$	$CS_{Cont1:6}$, $CS_{Cont2:7}$, $CS_{Cont3:8}$, $CS_{Cont4:9}$ CS_{Rand1}, CS_{Rand2}, CS_{Rand3}, CS_{Rand4}	4	96.9±1.8	92.4±2.5	94.7±1.3	94.7
		8	99.4±0.9	94.4±6.4	96.9±3.2	98.7
		16	100.0±0	98.9±1.0	99.4±0.5	100
		32	100.0±0	100.0±0	100.0±0	100

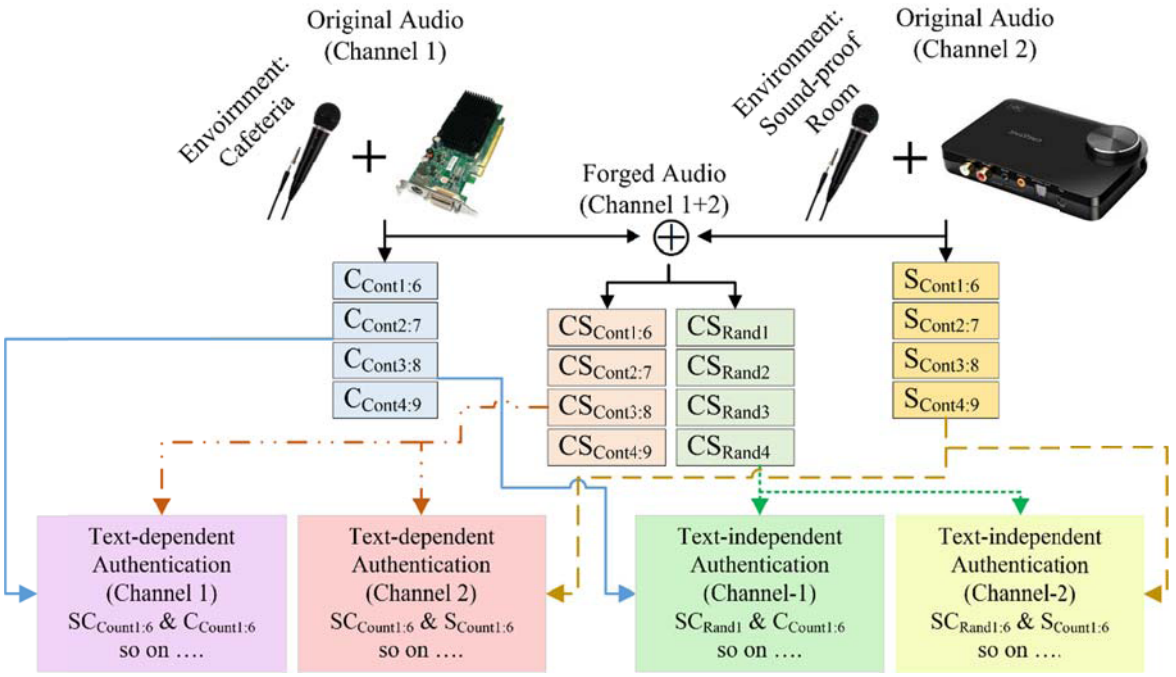

FIGURE 6. The setup for text-dependent and text-independent authentication.

and defined in Eqs. (12), (13) and (14). Different numbers of Gaussian mixtures (4, 8, 16 and 32) are used to perform the experiments. In addition, 10-folds cross validation is used in each experiment. All performance metrics are calculated for each fold. However, averaged results with a standard deviation (STD) are presented. From Table 2, it is evident that the accuracy of the system increased by increasing the number of GMM. This indicates that the forged and original audio modeled perfectly as the number of GMM increased. Moreover, the standard deviation of different folds also decreased when the number of GMM increased. The maximum SEN, SPE and ACC are achieved with 32 Gaussian mixtures, and they are 100%. STD is zero, which shows that the result is 100% for each fold.

In the second category, the classification of the different environments is performed. The environments of CDMB, SDME and forged audio are the cafeteria, sound-proof room, and combination of cafeteria and sound-proof room (Cafeteria+Room), respectively. In Table 3, classification accuracy is provided for the original audio of channel 1 (CDMB), channel 2 (SDME) and forged audio generated by merging both. The best accuracy for CDMB is 99.2%±2.6, for SDME it is 99.0%±2.1, and for the forged audio the accuracy is 99.2%±2.6. These results clearly indicate that the proposed system performed well in classifying different environments. It seems that the accuracy 99.2%+2.6 exceeds 100%. This situation occurs when average accuracy of folds is close to 100%, and some folds have accuracy away from the average.

In the third category, text-dependent authentication is performed. The training and testing of the proposed system is done by using the audio of the same text in these experiments, but the speakers are different. The speakers used in the training phase are not used during the testing of the system. The system authenticates the audio by comparing it with the acoustic models generated by using the different number of Gaussian mixtures. As shown in Fig. 6, text-dependent authentication is done with both channels (CDMB and SDME), one by one. The results are listed in Table 4. The maximum obtained accuracy for channels 1 and 2 is 100%±0.

TABLE 3. Classification of different environments: cafeteria, sound-proof room and cafeteria+room.

GMM	ACC±STD		
	CDMB	SDME	Forged
4	95.5±6.2	92.1±9.5	94.3±3.5
8	97.4±5.7	96.8±5.2	97.2±4.3
16	98.3±3.5	98.4±2.5	99.2±2.6
32	99.2±2.6	99.0±2.1	99.2±2.6

TABLE 4. Results for text-dependent authentication.

Normal Samples (Channel 1)	Forged Samples	GMM	ACC±STD	AUC	Normal Samples (Channel 2)	Forged Samples	GMM	ACC±STD	AUC
$C_{Cont1:6}$	$CS_{Cont1:6}$	4	95.7±5.8	95.9	$S_{Cont1:6}$	$CS_{Cont1:6}$	4	94.7±6.1	96.6
		8	98.7±2.8	99.1			8	98.9±2.4	99.9
		16	100±0	100			16	99.3±2.1	100
		32	100±0	100			32	99.3±2.1	100
$C_{Cont2:7}$	$CS_{Cont2:7}$	4	92.5±7.5	93.9	$S_{Cont2:7}$	$CS_{Cont2:7}$	4	95.0±5.1	95.6
		8	98.7±2.9	98.3			8	99.4±1.9	100
		16	100±0	100			16	98.9±2.4	100
		32	100±0	100			32	99.4±1.8	100
$C_{Cont3:8}$	$CS_{Cont3:8}$	4	94.3±3.6	94.0	$S_{Cont3:8}$	$CS_{Cont3:8}$	4	94.4±4.9	93.20
		8	99.1±1.9	98.1			8	98.2±2.4	100
		16	100±0	100			16	99.3±1.5	100
		32	100±0	100			32	99.6±1.1	100
$C_{Cont4:9}$	$CS_{Cont4:9}$	4	97.2±4.4	98.4	$S_{Cont4:9}$	$CS_{Cont4:9}$	4	93.6±7.7	94.4
		8	99.5±1.6	99.5			8	98.9±2.4	99.7
		16	100±0	100			16	100±0	100
		32	100±0	100			32	100±0	100

Furthermore, text-independent authentication is also conducted. Different text from the original and tampered audio is used to train and test the system. In these experiments, speakers as well as audio text are unknown to the system during the testing phase. Text-independent experiments are also performed for both channels, one by one. The obtained results are shown in Table 5. The best obtained accuracy for channel 1 is 100%±0 and for channel 2 is 99.5%±1.5.

In all experiments, the duration of audio is ∼5 seconds. Almost 100% accuracy is obtained to classify the original and forged audio in all categories of experiments. In each experiment, the speakers used to train the system are not used to test the system.

IV. DISCUSSION

By applying FT on the windowed blocks of an audio, a spectrum is obtained. The spectrum provides the energy information for each frequency component. Moreover, the spectrum is further processed by applying the principles of human psychoacoustics. The processed spectrum is our calculated feature vectors, and the proposed automatic authentication system is based on this. The processed spectrum of the digits, 1 and 2, for the three different environments is plotted in Fig. 7. The first environment is a cafeteria, and the audio is original; its spectrum is depicted in Fig. 7 (a). The second environment is a sound-proof room, and audio is original; its spectrum is shown in Fig. 7 (b). The third environment is a combination of a cafeteria and sound-proof room, and it is forged audio; its spectrum is plotted in Fig. 7(c).

The plotted spectrum shows the energy contours for digits 1 and 2. In the contours, red represents the high-energy regions, while blue signifies the lower-energy regions. A color bar is provided with each spectrum, and this is relative. For the original audio of channel 1, the energy

TABLE 5. Results for text-independent authentication.

Normal Samples (Channel 1)	Forged Samples	GMM	ACC±STD	AUC	Normal Samples (Channel 2)	Forged Samples	GMM	ACC±STD	AUC
Training $C_{Cont1:6}$ Testing $C_{Cont2:7}$	CS_{Rand1} CS_{Rand2}	4	94.2±5.3	94.2	Training $S_{Cont1:6}$ Testing $S_{Cont2:7}$	CS_{Rand1} CS_{Rand2}	4	96.64±4.6	96.5
		8	98.92±2.3	99.9			8	98.81±2.5	99.9
		16	100±0	100			16	99.33±2.1	100
		32	100±0	100			32	99.33±2.1	100
Training $C_{Cont3:8}$ Testing $C_{Cont4:9}$	CS_{Rand3} CS_{Rand4}	4	95.29±6.4	94.8	Training $S_{Cont3:8}$ Testing $S_{Cont4:9}$	CS_{Rand3} CS_{Rand4}	4	95.0±6.8	95.6
		8	99.33±2.1	99.2			8	95.5±5.1	99.6
		16	100±0	100			16	99.5±1.5	100
		32	100±0	100			32	99.5±1.5	100

TABLE 6. A comparison of the proposed system with existing studies.

Method	Comparison
Proposed System in this study (AAHP)	Classification of environments: 99.2%±2.6 Audio authentication for all original vs. all tempered audio: 100%±0 Audio authentication for text-dependent case: 100%±0 Audio authentication for text-independent case: 100%±0 (In all scenarios, the speakers used for testing were unknown to the system) (Audio is tempered splicing)
Chen et al. [21] 2016	89.50% (the best accuracy for forged audio) False positive rate: 13.74% False negative rate: 10.56% (Five different parameters need to be tuned) (Audio is tempered by deleting, inserting, substituting and splicing)
Zhao et al. [20] 2016	96.86±2.3% (over all accuracy) (Audio is tempered by splicing)

components lie in the range 0–5, while for the original audio of channel 2, the energy components lie in the range 0–12. After merging the two channels, the energy components of the forged audio range from 0 to 7. The digits 1 and 2 in the forged audio belong to channel 1 and channel 2, respectively. It can be observed from Fig. 7(c) that the energy contours and range of energy components are different from those of channels 1 and 2. The reason is that the forged audio contains audio parts from both channels, and the energy contour varies channel to channel according to the energy presented in the audio.

In this study, 24 band-pass filters are used. Therefore, the dimension of features for each divided block of an audio is 24. The interpretation of such high dimensional data is impossible by the human mind. Hence, a machine learning algorithm is used to make the automatic decision to differentiate between original and tampered audio.

In a recent study conducted by Chen et al. [21], audio is tampered by deleting, inserting, substituting and splicing. However, these operations change the audio significantly and someone can guess the forgery by listening to the tampered audio. In the study, no subjective evaluation is performed. It cannot be ignored that 80% or 90% of the forged samples can be detected truly by a human judge through visualization and hearing, and therefore an accuracy around 90% became an easy task. In another recent study conducted by Zhao et al. [12], the audio is forged by splicing, but deletion, insertion and substitution are not performed. In this study, subjective evaluation is also not performed. Despite these facts, a comparison of the proposed system with these studies is provided in Table 6.

In this study, an approach to generate the forged audio is also presented. The forged audio is generated with a great care so that a human judge cannot determine whether the

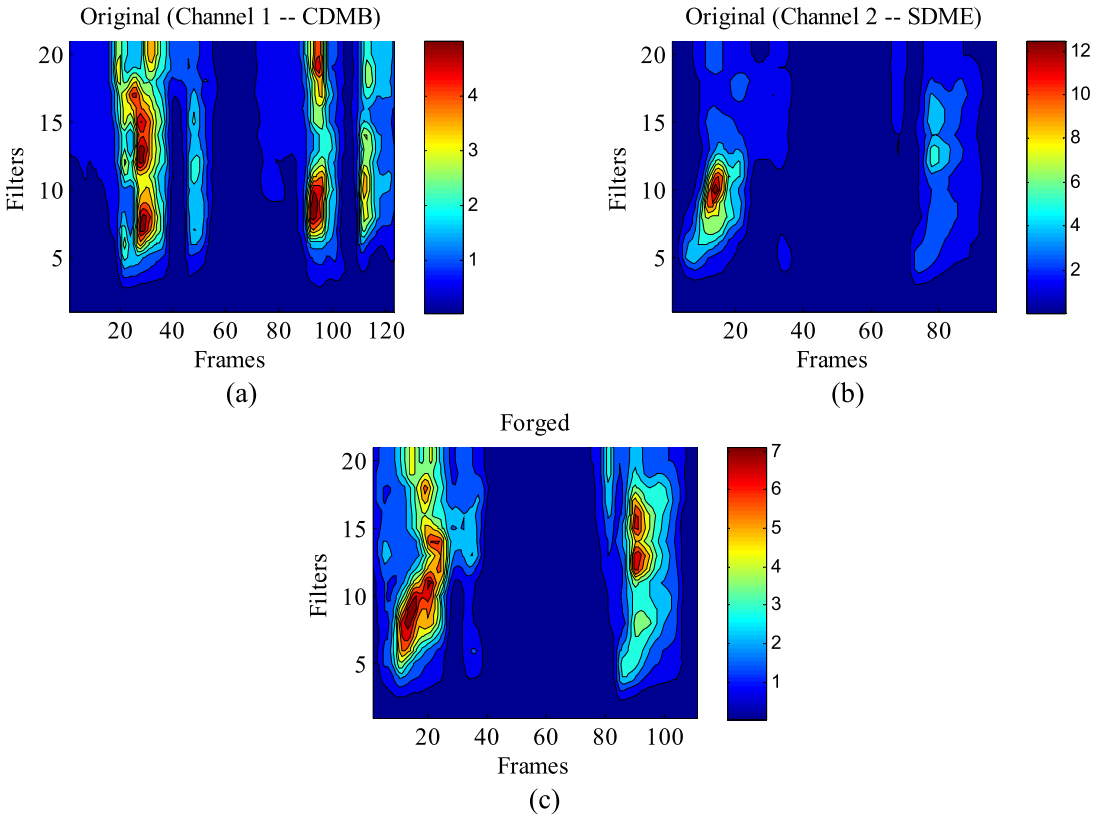

FIGURE 7. Energy contours for digits 1 and 2 in different spectrums (a) the original audio of channel 1 (b) the original audio of channel 2 (c) the forged audio (a combination of channels 1 and 2).

audio is original or tampered. The best obtained accuracy for the authentication of audio from the subjective evaluation is approximately 55%. Such accuracy confirms that the quality of the generated audio is excellent and cannot be judged by listening or visualizing. The accuracy of the proposed automatic audio authentication system is 45%, better than the best human judge.

V. CONCLUSION

This paper proposed an automatic audio authentication system based on three human psychoacoustic principles. These principles are applied to original and forged audio to obtain the feature vectors, and automatic authentication is performed by using the GMM. The proposed system provides 100% accuracy for the detection of forged and audio in both channels. The channels have the same recording microphone but different recording environments. Moreover, an accuracy of 99% is achieved for the classification of the three different environments. In automatic systems based on supervised learning, the audio text is vital. Therefore, both the text-dependent and the text-independent evaluation of the proposed system is performed. The maximum obtained accuracy is 100%. In all experiments, the speakers used to train and test the system are different (i.e., speaker-independent) and the obtained results are reliable, accurate and significantly outperform the subjective evaluation. The lower accuracy in the subjective evaluation also confirms that the forged audios are generated so sophisticatedly that human evaluators are unable to detect the forgery.

REFERENCES

[1] B. B. Zhu, M. D. Swanson, and A. H. Tewfik, "When seeing isn't believing [multimedia authentication technologies]," *IEEE Signal Process. Mag.*, vol. 21, no. 2, pp. 40–49, Mar. 2004.
[2] A. Piva, "An overview on image forensics," *ISRN Signal Process.*, vol. 2013, p. 22, Jan. 2013.
[3] A. Haouzia and R. Noumeir, "Methods for image authentication: A survey," *Multimedia Tools Appl.*, vol. 39, pp. 1–46, Aug. 2008.
[4] K. Mokhtarian and M. Hefeeda, "Authentication of scalable video streams with low communication overhead," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 730–742, Nov. 2010.
[5] S. Gupta, S. Cho, and C. C. J. Kuo, "Current developments and future trends in audio authentication," *IEEE Multimedia*, vol. 19, no. 1, pp. 50–59, Jan. 2012.
[6] R. Yang, Y.-Q. Shi, and J. Huang, "Defeating fake-quality MP3," presented at the Proc. 11th ACM Workshop Multimedia Secur., Princeton, NJ, USA, 2009.
[7] Q. Yan, R. Yang, and J. Huang, "Copy-move detection of audio recording with pitch similarity," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 1782–1786.
[8] X. Pan, X. Zhang, and S. Lyu, "Detecting splicing in digital audios using local noise level estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 1841–1844.
[9] A. J. Cooper, "Detecting butt-spliced edits in forensic digital audio recordings," in *Proc. 39th Int. Conf., Audio Forensics, Pract. Challenges*, 2010.
[10] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques-A review, and recent developments," *Signal Process.*, vol. 89, pp. 1489–1500, Aug. 2009.

[11] R. C. Maher, "Overview of audio forensics," in *Intelligent Multimedia Analysis for Security Applications*, H. T. Sencar, S. Velastin, N. Nikolaidis, and S. Lian, Eds. Berlin, Germany: Springer, 2010, pp. 127–144.

[12] B. E. Koenig and D. S. Lacey, "Forensic authentication of digital audio recordings," *J. Audio Eng. Soc.*, vol. 57, pp. 662–695, Sep. 2009.

[13] Audacity Team. (2016). *Audacity(R): Free Audio Editor and Recorder. Version 2.1.2 Retrieved*. [Online]. Available: http://www.audacityteam.org/

[14] GoldWave Inc. (2016). *GoldWave: Digital Audio Editing Software. Version 6.24 Retrived on November 25, 2016 From*. [Online]. Available: https://www.goldwave.com/goldwave.php

[15] C. Kraetzer, A. Oermann, J. Dittmann, and A. Lang, "Digital audio forensics: A first practical evaluation on microphone and environment classification," presented at the Proc. 9th Workshop Multimedia Secur., Dallas, TX, USA, Dec. 2007.

[16] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected MPEG-7 audio features and Mel-frequency cepstral coefficients," in *Proc. 5th Int. Conf. Digit. Telecommun.*, Jun. 2010, pp. 11–16.

[17] M. Huijbregtse and Z. Geradts, "Using the ENF criterion for determining the time of recording of short digital audio recordings," in *Computational Forensics: Third International Workshop*, Z. J. M. H. Geradts, K. Y. Franke, and C. J. Veenman, Eds. Berlin, Germany: Springer, 2009, pp. 116–124.

[18] D. P. Nicolalde and J. A. Apolinario, "Evaluating digital audio authenticity with spectral distances and ENF phase change," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Jan. 2009, pp. 1417–1420.

[19] D. P. N. Rodríguez, J. A. Apolinário, and L. W. P. Biscainho, "Audio authenticity: Detecting ENF discontinuity with high precision phase analysis," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 534–543, Sep. 2010.

[20] H. Zhao, Y. Chen, R. Wang, and H. Malik, "Audio splicing detection and localization using environmental signature," *Multimedia Tools Appl.*, pp. 1–31, Jul. 2016.

[21] J. Chen, S. Xiang, H. Huang, and W. Liu, "Detecting and locating digital audio forgeries based on singularity analysis with wavelet packet," *Multimedia Tools Appl.*, vol. 75, pp. 2303–2325, Jul. 2016.

[22] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer-Verlag, 2006.

[23] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, pp. 91–108, Aug. 1995.

[24] M. Alsulaiman, G. Muhammad, B. Abdelkader, A. Mahmood, and Z. Ali, *King Saud University Arabic Speech Database LDC2014S02*. Philadelphia, PA, USA: Linguistic Data Consortium, 2014.

[25] M. M. Alsulaiman, G. Muhammad, M. A. Bencherif, A. Mahmood, and Z. Ali, "King Saud University Arabic speech database," *Information*, vol. 16, pp. 4231–4253, Feb. 2013.

[26] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, "KSU Speech Database: Text Selection, Recording and Verification," in *Proc. Modeling Symp. (EMS)*, 2013, pp. 237–242.

[27] Y. Lin and W. H. Abdulla, "Principles of psychoacoustics," in *Audio Watermark: A Comprehensive Foundation Using MATLAB*. Cham, Switzerland: Springer, 2015, pp. 15–49.

[28] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[29] Z. Ali, M. Alsulaiman, G. Muhammad, I. Elamvazuthi, and T. A. Mesallam, "Vocal fold disorder detection based on continuous speech by using MFCC and GMM," in *Proc. GCC Conf. Exhibit. (GCC)*, Nov. 2013, pp. 292–297.

[30] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Amer.*, vol. 33, p. 248, Jan. 1961.

[31] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[32] S. S. Stevens, "On the psychophysical law," *Psychol. Rev.*, vol. 64, pp. 81–153, May 1957.

[33] J. Yang *et al.*, "Video compressive sensing using Gaussian mixture models," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4863–4878, Nov. 2014.

[34] J. I. Godino-Llorente, P. Gomez-Vilda, and M. Blanco-Velasco, "Dimensionality reduction of a pathological voice quality assessment system based on Gaussian mixture models and short-term cepstral parameters," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 10, pp. 1943–1953, Oct. 2006.

[35] T. H. Falk and W.-Y. Chan, "Nonintrusive speech quality estimation using Gaussian mixture models," *IEEE Signal Process. Lett.*, vol. 13, no. 2, pp. 108–111, Feb. 2006.

[36] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Rev.*, vol. 26, no. 2, pp. 195–239, 1984.

ZULFIQAR ALI received the master's degree in computational mathematics from the University of the Punjab, Lahore, Pakistan, and the master's degree in computer science from the University of Engineering and Technology, Lahore, with the specialization in system engineering. Since 2010, he has been a full-time Researcher with the Digital Speech Processing Group, Department of Computer Engineering, King Saud University, Saudi Arabia. He is currently a member of the Center for Intelligent Signal and Imaging Research, Universiti Teknologi PETRONAS, Malaysia. His research interests include speech and language processing, medical signal processing, privacy and security in healthcare, multimedia forensics, and computer-aided pronunciation training systems.

MUHAMMAD IMRAN has been an Assistant Professor with the College of Computer and Information Sciences, King Saud University, since 2011. He is currently a Visiting Scientist with Iowa State University, USA. He has authored number of high quality research papers in refereed international conferences and journals. His research is financially supported by several grants. His research interest includes mobile ad hoc and sensor networks, WBANs, M2M, IoT, SDN, fault tolerant computing, and security and privacy. He received a number of awards, such as the Asia Pacific Advanced Network fellowship. He served or is serving as a Guest Editor of the *IEEE Communications Magazine*, *Computer Networks*, *MDPI Sensors*, *International Journal of Distributed Sensor Networks*, the *Journal of Internet Technology*, and the *International Journal of Autonomous and Adaptive Communications Systems*. He has been involved in over 50 conferences and workshops in various capacities, such as a Chair, a Co-Chair, and a Technical Program Committee Member. These include the IEEE ICC, the Globecom, the AINA, the LCN, the IWCMC, the IFIP WWIC, and the BWCCA. Recently, European Alliance for Innovation (EAI) has appointed him as a Co-Editor-in-Chief of the *EAI Transactions on Pervasive Health and Technology*. He also serves as an Associate Editor of the IEEE Access, the *IEEE Communications Magazine*, the *Wireless Communication and Mobile Computing Journal*, the *Ad Hoc & Sensor Wireless Networks Journal*, *IET Wireless Sensor Systems*, the *International Journal of Autonomous and Adaptive Communication Systems*, and the *International Journal of Information Technology and Electrical Engineering*.

MANSOUR ALSULAIMAN received the Ph.D. degree from Iowa State University, USA, in 1987. Since 1988, he has been with the Computer Engineering Department, King Saud University, Riyadh, Saudi Arabia, where he is currently a Professor with the Department of Computer Engineering. His research areas include automatic speech/speaker recognition, automatic voice pathology assessment systems, computer-aided pronunciation training system, and robotics. He was the Editor-in-Chief of the *King Saud University Journal Computer and Information Systems*.